Eur J Psychol Educ
DOI 10.1007/s10212-014-0213-x

# Interpreting histograms. As easy as it seems?

**Stephanie Lem · Patrick Onghena · Lieven Verschaffel · Wim Van Dooren**

Received: 28 October 2013 / Revised: 2 January 2014 / Accepted: 24 January 2014


**Abstract** Histograms are widely used, but recent studies have shown that they are not as easy to interpret as it might seem. In this article, we report on three studies on the interpretation of histograms in which we investigated, namely, (1) whether the misinterpretation by university students can be considered to be the result of heuristic reasoning, (2) whether we could influence performance by stimulating or hindering the analytic processing of histograms, and (3) whether experts still show signs of this heuristic misinterpretation. We found that both university students and experts show signs of incorrect heuristic reasoning when comparing the mean of two data sets presented by histograms. Stimulating or hindering analytic processing did not affect performance. These indications of heuristic reasoning and the impossibility to affect this analytic processing suggest that the incorrect heuristic misinterpretation of histograms is very persistent. Implications for theory and methodology, scientific and educational practice are discussed.

**Keywords** Histograms · Misinterpretations · Experts · Dual-process theories · Heuristic reasoning

Histograms are omnipresent not only in research and education but also in popular media. Despite their omnipresence, various misinterpretations have been reported. In this article, we discuss three studies on one specific misinterpretation of histograms, using a dual-processing perspective. In the following sections we first give an overview of misinterpretations of histograms that have been documented in the literature, with special attention for the misinterpretation that we will study in this article, namely the interpretation of the height of the bars in a histogram as an indicator of the mean. Next, we introduce dual-process theories, which provide a theoretical and methodological basis to characterize these misinterpretations as the result of heuristic processing. Thereafter, we describe our research goals, and report on the

S. Lem (✉) · L. Verschaffel · W. Van Dooren
Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2, PO Box 3773, 3000 Leuven, Belgium
e-mail: Stephanie.Lem@ppw.kuleuven.be

P. Onghena
Methodology of Educational Sciences Research Group, KU Leuven, Dekenstraat 2, PO Box 3700, 3000 Leuven, Belgium

Published online: 08 February 2014

three conducted studies. Finally, we discuss the results, discuss the implications for theory and methodology, and for scientific and educational practice, and we formulate a conclusion.

## The misinterpretation of histograms

Various misinterpretations of histograms are described in the research literature. First, students mistakenly use height differences between individual bars of histograms as an indicator of variation, instead of using the range and overall shape of the figure (Baker et al. 2002; Cooper and Shore 2008; Lem et al. 2012. Second, students confuse histograms with bar graphs, thinking that the height of a bar in a histogram represents the value of that bar instead of the frequency or proportion (Baker et al. 2002; delMas, R et al. 2005). Third, they sometimes interpret the horizontal axis of a histogram as a time scale (delMas, R et al. 2005). Fourth, students misinterpret the classes in grouped histograms, thinking for example that each bar only represents the starting value. This creates difficulties with reading off frequencies of groups or values (delMas, R et al. 2005). A final misinterpretation—which will be the focus of the current article—is that students think that when one of two histograms has higher bars, it automatically also has a higher mean (Watson and Moritz 1998; Lem et al. 2012). Figure 1 illustrates this misinterpretation: Both data sets shown in the figure have the same mean (i.e., 6), but students think that the higher bars in the right histogram imply that the mean in that data set is also larger, while this is not the case.

From now on, we will refer to this last misinterpretation as the "height misinterpretation". Lem et al. (2012) claimed that the height interpretation exists due to the fact that the way in which people tend to interpret graphs is not in line with the design of histograms. This claim was based on the graph design principles proposed by Tversky (1997). According to Tversky, the way we interact with the world influences how we interpret graphs. One of these principles poses the perceptual preference of the vertical above the horizontal dimension.

Gravity is correlated with vertical, and people are oriented vertically. The vertical axis of the world has a natural asymmetry, the ground and the sky, whereas the horizontal axis of the world does not. The dominance of the vertical over the horizontal is reflected in the dominance of columns over rows. It is more usual and more natural to make a vertical list than a horizontal one. Similarly, bar charts typically contain vertical columns. (Tversky 1997, p. 120–121).

According to Tversky (1997), also in interpreting graphs, we tend to focus more on the vertical than on the horizontal dimension. Very often, this is useful, for instance to read off the number of observations with a specific value). However, when we compare the mean of two histograms, the horizontal dimension is at least as important as the vertical dimension. If the

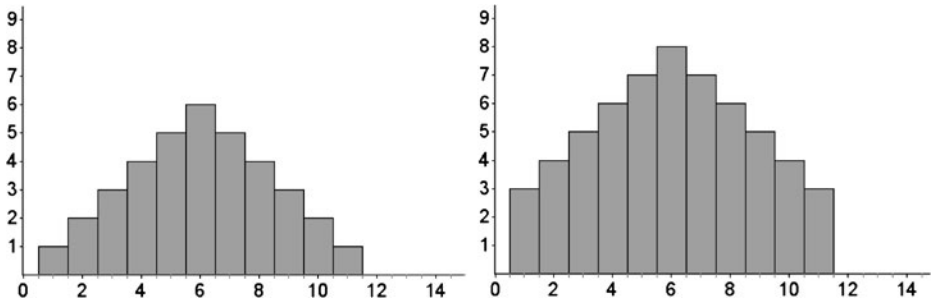



**Fig. 1** Students think that the *right histogram*, with higher bars, represents a data distribution with a higher mean than the *left histogram*, with lower bars

bars of a given histogram become higher, this is an indication that more observations were included,[1] but the mean of the data set is not affected. However, if the bars of a histogram are situated more to the right of the X-axis, the mean of the data set will be larger, all other things being equal. If students naturally tend to pay more attention to the vertical dimension than to the horizontal dimension, this may explain why they misinterpret the difference in height of the bars of two histograms as representing the difference in means in these two histograms (Lem et al. 2012). One of the main goals of this article is to investigate whether the height misinterpretation can be considered intuitive or heuristic in nature. In order to do this, we used a methodology often applied in research using the dual-process theoretical framework.

## Dual-process theories

It is quite surprising that students have so many difficulties interpreting histograms. Not only are they confronted with histograms from a primary school age on, both in school and in popular media, histograms also seem to be among the more easily interpretable representations in statistics (as compared to, for instance, box plots; Lem et al. 2013a. How can we then understand the occurrence of the height misinterpretation? The dual-process theoretical framework may provide an answer to this question, because it was developed to help understand why people who possess the required knowledge to solve a certain problem, may nevertheless fail to give a correct answer. The dual-process framework assumes that this is due to the impact of heuristic processing of certain salient but not necessarily relevant problem characteristics (e.g., Kahneman and Tversky 1972). This framework has recently been successfully applied to explain erroneous answers to mathematical problems, like proportional word problems and rational number tasks (e.g., Gillard et al. 2009; Vamvakoussi et al. 2013 and also to explain the misinterpretation of box plots (Lem et al. 2013b).

According to dual-process theories of reasoning, by default one tends to use very fast reasoning processes—called heuristic or intuitive processes—when interpreting a situation or task, and only in some cases (for instance when very attentive) one will also employ slower and more effortful analytic reasoning processes. Heuristic reasoning processes very frequently lead to the correct solution (the heuristic system probably originates and survives because of its relative effectiveness at high speed with low effort), but analytic reasoning is necessary for tasks where heuristic reasoning is not successful. According to the so-called extended heuristic–analytic model of St. Evans (2006), heuristic and analytic reasoning work in constant competition and interaction with each other. When confronted with a task, one will immediately start to construct the most plausible or relevant *default model*, based on automatically processed salient task features, the goal of the task, and background knowledge. Only after this initial heuristic processing, analytic reasoning *may* also be initiated, but whether this will happen depends on many factors such as general intelligence, time available, and task instructions. When analytic reasoning takes place, the default model's validity is evaluated (based on the amount of cognitive conflict experienced) and possibly modified before a final response is given.

An important consequence of Evans' model is that even when analytic thinking takes place, reasoning may still be based on—and therefore biased by—the default model and (possibly irrelevant) salient task features that were already processed at the very first confrontation with the task. So, heuristically processed task features may still interfere in the analytic stage of the

[1] This is of course only the case when the histogram represents absolute frequencies, not in the case of relative frequencies or cumulative frequencies. In this manuscript, we will only address frequency histograms.

reasoning process and have an important influence on the final outcome of the reasoning, in the form of a heuristic response. From this extended model, we can derive that even experts, who are able to correctly solve the mathematical tasks at hand, could still be influenced by intuitions or heuristic reasoning (e.g., Inglis and Simpson 2004), if not in their ultimate answer, at least in their solution process.

Applying this to histograms, we expected that the height of the bars would act as a salient task feature (Tversky 1997) that may in some cases have a negative effect on the reasoning process and its final outcome. When comparing the mean of two histograms, the almost immediately heuristically generated default model would be that the histogram with the higher bars also has the higher mean. Even when people know that they should take the horizontal dimension and overall shape of the histogram into account, and apply this knowledge analytically to the task at hand, the default model may place its mark on the outcome of the reasoning and lead to an incorrect response to the task. For those who succeed in giving the correct response to the task, we anticipate longer processing times because relatively time-consuming analytic reasoning processes are necessary to overcome the default model.

## Research goals

In a series of three experiments, we investigated the height misinterpretation of histograms, using a dual-processing approach. First, we wanted to know whether this misinterpretation can be characterized as an error that results from heuristic processing. Second, we wanted to test the effect of experimentally stimulating or hindering the analytic processing of histograms, in order to get insight in the strength of the heuristic reasoning error. Third, we tested whether this heuristic reasoning error is present not only in common users but also in expert users of histograms.

## Study 1. A heuristic error?

The goal of the first study was to test whether the height misinterpretation is indeed the result of heuristic processing. In order test this hypothesis, we used a common method from the dual-processes literature (e.g., Gillard et al. 2009; Lem et al. 2013b), by comparing accuracy rates and reaction times for two types of items: congruent items and incongruent items. Congruent items are items for which the heuristic response is correct, meaning that heuristic reasoning is sufficient to respond correctly. For incongruent items on the other hand, the heuristic response is not correct and slower analytic reasoning is necessary to find the correct response. This method hence uses the speed with which both processes are executed as an important parameter. Concerning accuracy, this method expects most correct responses for congruent items and least for incongruent items. Concerning reaction times, this method expects relatively long reaction times for correct responses to incongruent items, as slow analytic reasoning is necessary to find this correct response.

### Method

#### *Participants*

Participants were 40 first-year students of educational sciences. All participants had completed the same introductory statistics course several weeks before participation, covering histograms

and descriptive statistics among various other topics. In order to be sure that the participants indeed had the required knowledge to solve the items in the test, we asked them to interpret two histograms by stating the frequency of which several bars were observed. Participants scored, on average, 98.7 % on this test, so we assume that their knowledge of histograms was sufficient to solve the items in the experiment. Participants were engaged in the experiment as a part of course requirements.

*Materials*

Participants were presented with 40 histogram comparison items in random order. Each item consisted of two histograms representing fictitious exam results of two groups of students. Participants' task was to determine which of the two groups had the largest mean exam result. In all items, the only two elements being varied between both histograms were the height of the bars and/or their horizontal location. Assuming that the height of the bars would be processed heuristically while their horizontal location would require analytic processing, we constructed two major types of items: congruent and incongruent ones. In congruent items, a heuristic processing based on the height of the bars leads to the correct answering alternative, making analytic reasoning unnecessary. In incongruent items, heuristically processing the height of the bars leads to the incorrect alternative, so this heuristic process needs to be inhibited and the horizontal location of the bars needs to be processed analytically in order to obtain the correct response.

Manipulating the height of the bars as well as their horizontal location allowed to construct two types of congruent and three types of incongruent items (see Fig. 2): *Congruent equal items* present two identical histograms. For these items, the heuristic response that the two histograms have the same mean is correct. Also, in *congruent unequal items*, the heuristic response is the correct one, as the histogram with the higher mean also has higher bars. In *incongruent equal items*, the correct response is that the mean is the same in both histograms, while the heuristic response would be that the histogram with higher bars has a higher mean. In *incongruent inverse items*, the histogram with the bars placed more to the left has higher bars, leading to the heuristic response that this histogram has a higher mean. The correct response, however, is that the histogram with the bars placed more to the right has the highest mean, notwithstanding its lower bars. Finally, *incongruent unequal items* show two equally shaped histograms, in one of which the bas are placed more to the right. The heuristic response would be that both histograms have the same mean as the height of both histograms is the same, while the correct response is that the histogram with the bars placed more to the right has the largest mean.

*Procedure*

The histogram comparison test was administered in groups of 20 students in a computer class where each student worked individually on a computer. The 40 items (16 congruent, 24 incongruent) were provided in blocks of ten items each, followed by a break which participants could end by tapping the space bar. After the general introduction of the task, two sample items were provided without feedback. Students were told to work at their own pace and to try to respond correctly. All items were preceded by a fixation cross in the center of the screen which was presented for 500 ms. The items were presented in a semirandom order, with the following restrictions: (a) not more than three times the same item type on consecutive trials, (b) not more than three times the same heuristic response on consecutive trials, (c) not more than three times the same correct

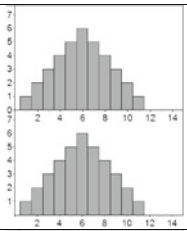
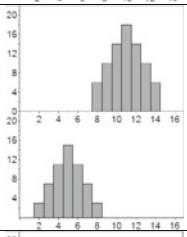
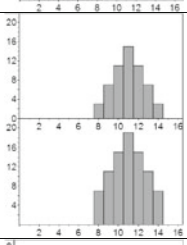
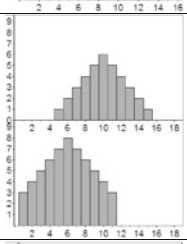
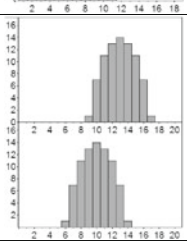
| | Correct response | Heuristic response | Example |
|---|---|---|---|
| Congruent equal | The same (same location and shape) | The same (same height) |  |
| Congruent unequal | Histogram that is located more to the right | Histogram with highest bars |  |
| Incongruent equal | The same (same location and shape) | Histogram with higher bars |  |
| Incongruent inverse | Histogram located more to the right | Histogram with highest bars |  |
| Incongruent unequal | Histogram located more to the right | The same (same height) |  |

**Fig. 2** Overview of the five item types used in Study 1. The task was to decide, for each pair of histograms representing the exam results of two groups of students, which group had the highest mean score

response on consecutive trials, and (d) not more than three times the same level of congruency on consecutive trials. Stickers were placed on the 9, 6, and 3 of the numerical keyboard, and participants were asked to press the 9 (with sticker "up") when their answer was "top histogram," 6 (with sticker "=") when the answer was "both the same," or 3 (with sticker "down") when their answer was "lower histogram." For each item, the reaction time and accuracy were logged.

## Predictions

With respect to accuracy, we expected congruent items to be solved better than incongruent items (Prediction 1), as heuristic reasoning would be sufficient to obtain the correct response for congruent items, while analytic reasoning would be necessary for correctly solving the incongruent items. Concerning reaction time, we anticipated that correct responses to congruent items would be given faster than correct responses to incongruent items (Prediction 2). This is because analytic reasoning, which is necessary to find the correct response to incongruent items, takes longer than heuristic reasoning, which suffices for finding the correct response to congruent items.

## Results

Outliers with respect to reaction time—calculated per item type—were removed first. This resulted in the deletion of the 16 (1.0 %) responses with a reaction time more than 2.5SD from the mean reaction time of the corresponding item type.

### *Accuracy*

We used a generalized linear mixed model with accuracy as dependent variable and congruency as independent variable. We found a main effect of congruency, $F(1,1543)=205.00$, $p<0.001$, $OR=18.04$. The higher accuracy for congruent items (92.9 %) than for incongruent items (60.6 %) confirmed Prediction 1. In addition, we found that 92.5 % of all incorrect responses to incongruent items were the heuristic response.

A closer look at the accuracy rates per incongruent item type showed an interesting pattern (see Fig. 3): The mean accuracy for incongruent equal items was only about half (41.8 %) of the accuracy for incongruent unequal items (77.7 %), while the accuracy for incongruent inverse items was in between (62.3 %). A generalized linear mixed model on the incongruent items, with accuracy as dependent variable and item type as independent variable, showed that this difference was statistically significant, $F(2,908)=60.82$, $p<0.001$, $OR=0.04$–$0.21$. The different accuracy for the three incongruent item types can be explained by studying these

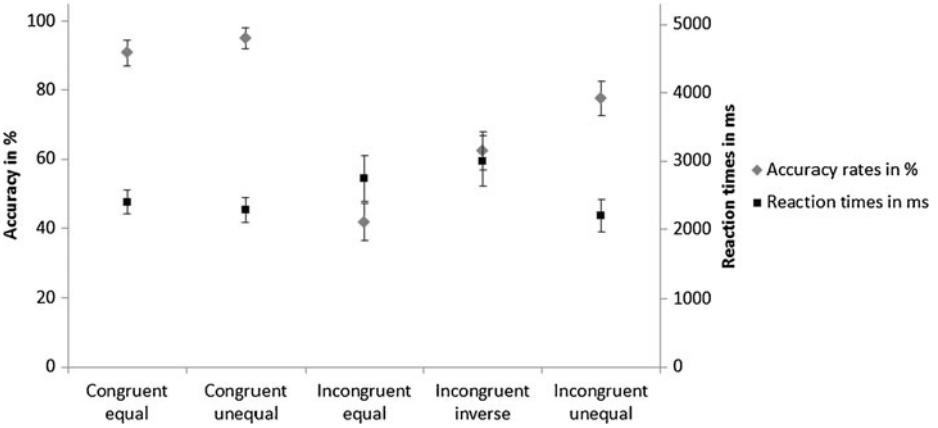



**Fig. 3** Accuracy rates and reaction times per item type in Study 1. Reaction times are for the correct responses only. Error bars for the accuracy rates show Clopper–Pearson confidence intervals. Error bars for the reaction times show the 95 % confidence intervals

three item types more closely. The incongruent unequal items elicited most correct responses, since there was no height difference that could distract students, while there was a difference in location that could lead to the correct conclusion (see Fig. 2). So students used the only available difference between the histograms to compare the means, and this difference was a relevant one. In the incongruent equal items, which elicited fewest correct responses, the only difference between the two histograms was the height of the bars. If students relied on this difference between the histograms in order to compare the means, this time they arrived at the erroneous heuristic response. In the incongruent inverse items, finally, which had a medium accuracy rate, both the position and the height differed, making it attractive to focus on the height difference, although analytic reasoning might interfere and possibly lead to an ultimate correct response.

*Reaction time*

To analyze students' reaction time on the correctly solved items, we used a linear mixed model with the log reaction time as dependent variable and congruency as independent variable. We found again a main effect of congruency, $F(1,1124)=12.09$, $p<0.001$, $d=0.13$.[2] The difference in reaction time between both levels of congruency confirmed Prediction 2, with slightly but statistically significant faster responses to congruent items (2,342 ms, $SD=1{,}518$) than to incongruent items (2,599 ms, $SD=2{,}239$).

A closer look at the reaction times per item type for the incongruent items only revealed a main effect of item type as well, $F(2,536)=30.60$, $p<0.001$, $d=0.11$–$0.35$ (see Fig. 3). Incongruent unequal items were solved fastest (2,198 ms, $SD=1{,}930$), followed by incongruent equal (2,746 ms, $SD=1{,}977$) and incongruent inverse items (3,002 ms, $SD=2{,}653$). This pattern was somewhat unexpected, as the very well-solved incongruent unequal items were also solved with nearly the same speed as congruent items.

## Study 2. How strong is the height heuristic?

In Study 1, we found evidence for the existence of an incorrect reasoning process that was heuristic in nature, making students compare the heights of the bars in two histograms instead of taking into account the distribution as a whole when comparing the histograms with respect to their means. We can hence indeed call the height misinterpretation a "height heuristic." In a second study, we tried to get deeper insight into the strength of this heuristic reasoning error by working with two new conditions that aimed at stimulating or hindering analytic reasoning. In a first condition, the analytic condition, we tried to maximize the occurrence of analytic reasoning by providing unlimited solving time and warning students about the possible fallaciousness of graphs, using an example of a misleading line graph, at the beginning of the experiment. In the second condition, the heuristic condition, we tried to minimize the interference of analytic reasoning by limiting the allowed response time so that time-consuming analytic processing could not successfully occur. Also, the warning that was given in the analytic condition was omitted. Comparison of the responses in these two conditions can give an indication of the strength of the height heuristic: Assuming that the heuristic would be relatively weak and susceptible to manipulation, students in the analytic condition should perform better than students in the heuristic condition. If, on the other hand, the condition

[2] Cohen's $d$ between 0.2 and 0.3 can be considered a small effect, around 0.5 the effect can be considered medium, and a Cohen's $d$ of 0.8 or higher can be considered to be large (Cohen, Manion, & Morrison 2007).

wherein analytic reasoning is stimulated would not lead to better performance than in the condition wherein it is prohibited, this would point at a very strong heuristic that is resistant to analytic reasoning.

## Method

### *Participants*

Participants were 74 first-year students of educational sciences. The students came from the same group as the ones from Study 1 and had hence all completed the same introductory statistics course several weeks before participation, covering histograms and descriptive statistics among various other topics. The students solved the same histogram items as the participants in Study 1 before taking the experiment, scoring, on average, 98.1 %. Participants engaged in the experiment as a part of course requirements. Students who already participated in Study 1 were excluded from this study.

### *Materials*

We used the same histogram comparison test as in Study 1. This time, however, we manipulated the allowed reaction time and the instructions. In the *heuristic condition* ($n$=39), students were told they had to respond within 4 s. In the *analytic condition* ($n$=35), students were told they could use as much time as needed. Additionally, at the beginning of the experiment, the students of the analytic condition were given an example exercise concerning a line graph to warn them about the fact that graphs can be misleading and that this could also be the case in the test they were about to take. In this example exercise, two graphs were shown in which the distance walked was presented for a girl who was hiking in the mountains. Students' task was to choose the graph in which the girl encountered a very steep climb after 50 min (see Fig. 4). In one graph a very steep line was shown after 50 min, while a much less steep line was shown at that point in the other graph. At first sight, one might think that the very steep line represents a steep climb, while, of course, the girl would cover much less distance during the steep climb, represented by a much less steep line in the other graph.

### *Procedure*

Just like in Study 1, students took the test in groups of maximum 20 students individually in a computer class. Groups of students were randomly assigned to either the heuristic or the analytic condition. We chose to not have different conditions in one group, as especially the

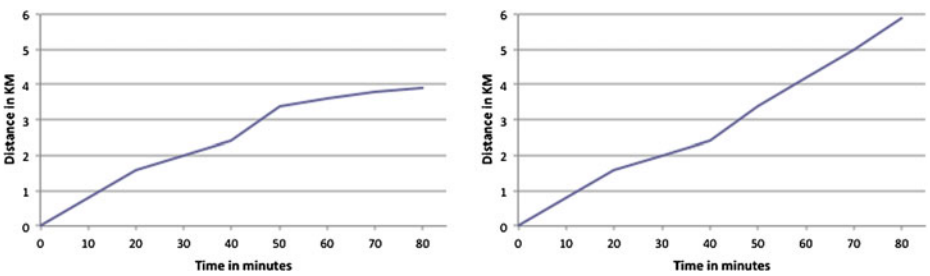



**Fig. 4** The line graphs used as a warning for students in the analytic condition

time constraint of students in the heuristic condition could also affect the response times of students in the analytic condition. The rest of the procedure was the same as in Study 1.

## Predictions

With respect to accuracy, we expected the same pattern of more correct responses to congruent items than to incongruent items as in Study 1 (Prediction 1). However, assuming that the impact of the heuristic would be susceptible to the stimulation of analytic processing, we also anticipated more correct responses in the analytic condition than in the heuristic condition, particularly for the incongruent items (Prediction 2). The reason for this second prediction is that participants in the heuristic condition would barely have the time to employ the analytic reasoning processes necessary to find the correct responses to the incongruent items. In the analytic condition, on the other hand, we anticipated higher accuracy rates as the students had enough time to reason analytically and were also extra motivated to do so by means of the warning.

With respect to reaction time, we predicted the same pattern of longer reaction times for correct responses to incongruent items than to congruent items as in Study 1 (Prediction 3).

## Results

Outliers with respect to reaction time were calculated per item type and condition. This resulted in the deletion of the 22 (0.8 %) responses with a reaction time more than 2.5SD from the mean reaction time of the corresponding item type.

### *Accuracy*

We found the expected pattern of congruent items being solved better than incongruent items, in both the heuristic condition, $F(1,1490)=165.60$, $p<0.001$, $OR=15.55$, and the analytic condition, $F(1,1343)=166.11$, $p<0.001$, $OR=23.72$. Accuracy was highest in the congruent items (94.8 % in the heuristic condition, 95.7 % in the analytic condition) and lowest in the incongruent items (63.6 % in the heuristic condition, 63.4 % in the analytic condition), confirming Prediction 1. However, no main effect of condition was found, $F(1,2835)=0.20$, $p=0.659$, which disconfirms Prediction 2. Like in Study 1, we found very high percentages of heuristic responses with incorrectly solved incongruent items: 93.7 % for the heuristic condition, and 94.2 % in the analytic condition.

In addition to testing our predictions, we checked for the same pattern of accuracy rates for the different incongruent items as we found in Study 1. We found the same effect, $F(2,1662)=138.02$, $p<0.001$, $OR=0.03–0.17$, and pattern of accuracy rates for the three incongruent item types as in Study 1 (see Figs. 5 and 6). Incongruent unequal items were solved best (83.0 % in the heuristic condition and 86.1 % in the analytic condition), incongruent inverse items followed (with 67.4 % and 61.2 % correct answers in the heuristic and analytic condition, respectively), and incongruent equal items were solved worst (with 40.1 % and 39.9 % correct answers in the heuristic and analytic condition, respectively).

### *Reaction time*

As in Study 1, we only analyzed the reaction time of the correct responses. A general linear mixed model with reaction time as dependent variable and congruency and condition as independent variables, together with their interaction effect, showed no main effect of

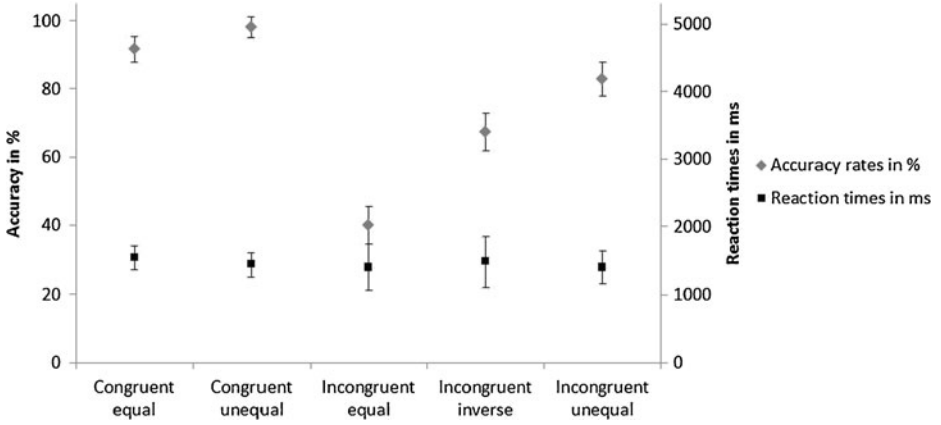



**Fig. 5** Accuracy rates and reaction times per item type for the heuristic condition in Study 2. Reaction times are for the correct responses only. Error bars for the accuracy rates show Clopper–Pearson confidence intervals. Error bars for the reaction times show the 95 % confidence intervals

congruency, $F(1,2134)=1.22$, $p=0.269$, which was different from our third prediction and from the results of Study 1. A main effect of condition was found, $F(1,2134)=83.34$, $p<0.001$, $d=0.89$, which was in line with the intended experimental manipulation: Reaction times for correct answers were much longer in the analytic condition (3,426 ms, $SD=3045$) than in the heuristic condition (1,462 ms, $SD=641$). There was no interaction effect of congruency and condition, $F(1,2134)=0.00$, $p=0.978$, suggesting that the effect of condition was the same for both levels of congruency.

In addition to testing our predictions, we checked for the pattern of reaction times for the different incongruent items as we found in Study 1. We again found an effect of item type on reaction time, $F(2,1021)=25.64$, $p<0.001$, $d=0.07$–$0.27$, following the same pattern as in Study 1. The reaction times in the heuristic condition did not differ significantly between the three types of incongruent items (1,408 ms for incongruent equal, 1,483 ms for incongruent inverse, and

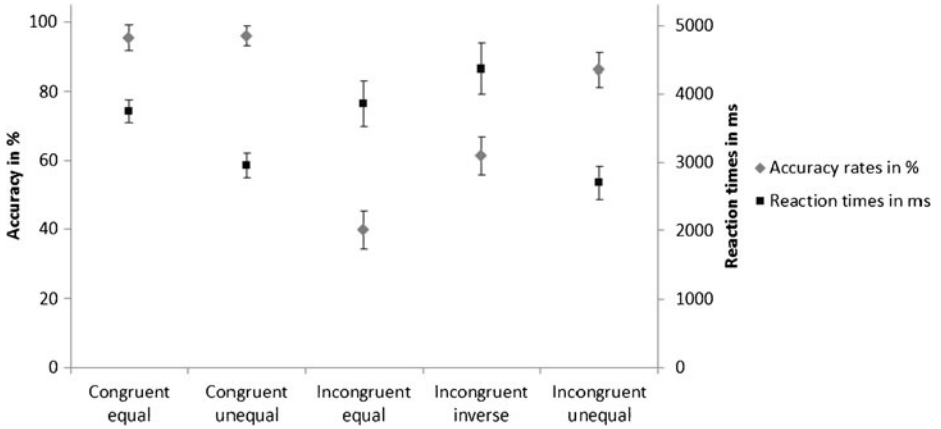



**Fig. 6** Accuracy rates and reaction times per item type for the analytic condition for Study 2. Reaction times are for the correct responses only. Error bars for the accuracy rates show Clopper–Pearson confidence intervals. Error bars for the reaction times show the 95 % confidence intervals

1,404 ms for incongruent unequal items; $F(2,540)=3.96$, $p=0.012$ ), but they did follow the same pattern as in Study 1. In the analytic condition we saw the same pattern as in Study 1 and in the heuristic condition, but much more pronounced, resulting in a statistically significant main effect of item type on reaction time: $F(2,479)=21.72$, $p<0.001$, $d=0.12$–$0.20$. Again, incongruent unequal items were solved fastest (2,699 ms)—even faster than the congruent items in this condition—the incongruent equal items took somewhat longer (3,865 ms), and the incongruent inverse items took longest (4,372 ms), as shown in Figs. 5 and 6.

## Study 3. Do experts also fall prey to the height heuristic?

Studies 1 and 2 showed that heuristic reasoning leads university students who were only recently taught about histograms, to the incorrect interpretation of the height of bars in histograms when comparing means (Studies 1 and 2), and that the heuristic is strong, as an experimental stimulation or hindering of analytic reasoning did not affect performance (Study 2). Given the strength of the heuristic in these novice users, we tested in Study 3 whether even expert users of histograms are susceptible to heuristic reasoning (Fig. 7).

Research on expertise in various domains has revealed that experts focus more on structural principles of a task, while novices rely more on surface features (Hardiman et al. 1989; Rabinowitz and Hogan 2002, 2008; Inglis and Alcock 2011). An example of experts' greater focus on structural principles in the domain of statistics can be found in the study of Rabinowitz and Hogan (2002). University students with varying levels of experience in statistics had to match various statistical problems to each other. While more experienced students focused on more structural features of the presented problems (e.g., type of test to be used to solve the problem), less experienced students tended to focus more on surface features (such as the narrative cover story of the problem).

Applying these findings from expert research to the interpretation of histograms, one would expect that—unlike students who received only some instruction in histograms—expert users of histograms are no longer hampered by the height heuristic, which relies on a superficial feature of histograms, namely, the height of the bars. Due to their extensive experience with this graphical representation, experts can be assumed to immediately look at the correct task features, i.e., the combination of the overall shape and the horizontal position of the bars of the histogram.

On the other hand, if experts would still be affected by the height heuristic, they should show the same effects of congruent and incongruent items on accuracy and/or reaction time as the students in Studies 1 and 2 (albeit perhaps to a lesser extent). So, if experts would still perform less accurately on incongruent than on congruent items, we would have to conclude that the height heuristic continues to play a significant role even in experts' reasoning processes about histograms. But even when their accuracy would remain unaffected and only their response times on correctly solved incongruent items would be longer than on correctly solved congruent items, this would be a clear indication that the height heuristic is still influential.

### Method

#### *Participants*

Participants were 40 students and staff who could be considered as expert users of histograms. We defined expert users of histograms as people who do not only have been instructed

extensively about histograms like the students in Studies 1 and 2 but also work with histograms on a regular basis. We found four groups of people who fit this definition: students of the master of statistics program ($n$=5), researchers in statistics ($n$=8), statistics professors ($n$=9), and researchers of more quantitatively oriented subfields of psychology, sociology, and educational sciences in which histograms are regularly used ($n$=18). Participants were recruited by e-mail and participated on a voluntary basis.

*Materials*

The materials used were exactly the same as the ones used in Study 1. This means that the experts solved the same 16 congruent items and 24 incongruent items and that their reaction times and responses were logged.

*Procedure*

The procedure followed was the same as in Study 1, with the only difference that the test was administered individually at laptop computers instead of groupwise in a PC room.

Predictions

If experts are still influenced by the height heuristic, this could be visible in the same two ways as with the students in Studies 1 and 2. First, accuracy for congruent items could be higher than for incongruent items (Prediction 1). Second, reaction times of correctly answered incongruent items would be longer than reaction times of correctly answered congruent items (Prediction 2).

Results

Outliers with respect to reaction time were calculated per item type and removed before analysis. This resulted in the deletion of the 31 (2.0 %) responses with a reaction time more than 2.5SD from the mean reaction time of the corresponding item type. Of the incorrectly solved incongruent items, 93.4 % was incorrect because the heuristic response was provided.

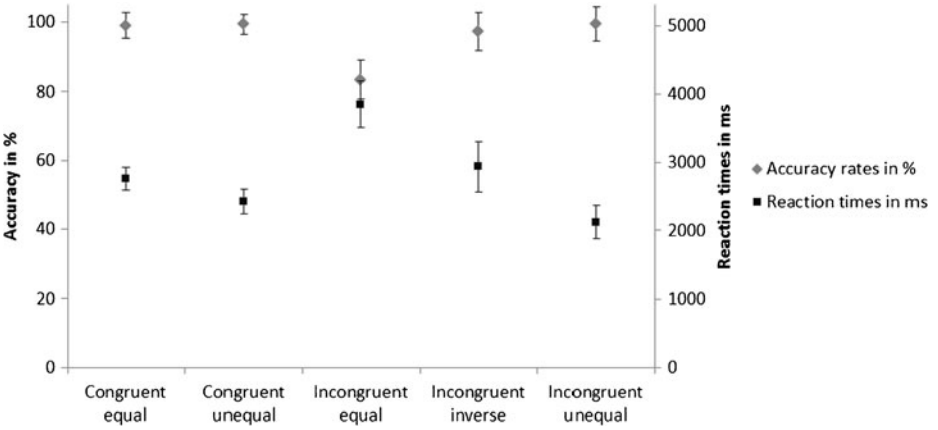



**Fig. 7** Accuracy rates and reaction times per item type in Study 3. Reaction times are for the correct responses only. Error bars for the accuracy rates show Clopper–Pearson confidence intervals. Error bars for the reaction times show the 95 % confidence intervals

*Accuracy*

A generalized linear mixed model with accuracy as dependent variable and congruency as independent variable showed a main effect of congruency on accuracy, $F(1,1489)=24.75$, $p<0.001$, OR=10.92. As in the two previous studies, accuracy was higher for congruent items (99.2 %) than for incongruent items (93.4 %). This effect of congruency on accuracy suggested that experts were, like students, still influenced by the height heuristic, confirming Prediction 1.

Because of the interesting pattern of accuracy rates for the different incongruent item types we found with the students in Studies 1 and 2, we also fitted a generalized linear mixed model to the incongruent items only, to test the effect of item type on accuracy, next to testing our predictions. We found a main effect of item type on accuracy, $F(2,876)=26.09$, $p<0.001$, $OR$= 0.01–0.19. The pattern was a bit different than for the students, as accuracy for incongruent unequal items (99.4 %) and incongruent inverse items (97.3 %) were almost the same and even generally at the same level as that of the congruent items, while accuracy to incongruent equal items was significantly lower (83.5 %), $F(1,800)=8.24$, $p<0.001$, $OR$=0.06–0.31. Note that this was also the most difficult item type for the students in Study 1 and Study 2.

*Reaction time*

A linear mixed model with reaction time as dependent variable and congruency as independent variable, performed on the correct responses only, showed a main effect of congruency on reaction time: $F(1,1423)=6.37$, $p=0.012$, Cohen's $d=0.12$. The effect was in the same direction as with the students in Studies 1 and 2, as correct responses to incongruent items (2,924 ms, $SD$=3,268) took longer than to congruent ones (2,593 ms, $SD$=2,122).

In addition to testing our predictions, we also checked for the same reaction time pattern for the different incongruent items as we found in Studies 1 and 2. As with the students, a main effect of item type on reaction time was found for the correct responses to incongruent items, $F(2,815)=53.75$, $p<0.001$, but the direction of the effect was somewhat different. As in Studies 1 and 2, incongruent unequal items were solved fastest (2,123 ms). Different, however, was that not the incongruent inverse (2,940 ms), but the incongruent equal items were solved slowest (3,854 ms). Incongruent equal items are items where only the height of the histograms differs and where the smallest accuracy rates were found. This could be interpreted as an indication that experts got confused when confronted with this item type, needing extra time to convince themselves of the correctness of their analytic reasoning, in which they did not always succeed. This is similar to the general conclusion that can be drawn concerning the different incongruent item types: the lower the accuracy rate, the longer the reaction times, suggesting more time to spend on analytic reasoning.

## Conclusion and discussion

In this series of three studies, we investigated the occurrence of one specific misinterpretation of histograms, namely, the height misinterpretation using a dual-processing perspective. In Study 1, we compared accuracy rates and reaction times for congruent and incongruent items to test whether the height misinterpretation of histograms could be characterized as the result of incorrect heuristic processing. In the second study, we tried to stimulate or hinder students' analytic reasoning in order to test the strength of the heuristic error. In the third and final study, we subjected expert users of histograms to the same experiment as in Study 1, to test whether even experts would be misled by the height heuristic, if not in their response patterns, at least

in their reaction times. Several conclusions can be drawn. First, students' display of the height misinterpretation can be explained by their heuristic processing of the height of the bars in histograms, as evidenced by both their accuracy rates and reaction times. Incongruent items, in which heuristic reasoning is insufficient, were solved worse than congruent items, which can be solved correctly by means of heuristic reasoning alone. Also, when incongruent items were solved correctly, this took longer than when congruent items were solved correctly, as more time-consuming analytic processing was necessary. Second, students who were forced to respond within a few seconds performed equally well as students who could use as much time as they wanted and were warned about the possible fallaciousness of graphical representations. This indicates that this heuristic reasoning is strong and that the outcome of the heuristic processing cannot easily be manipulated. Third, even experts displayed the height heuristic, again demonstrating the strength of the heuristic. Fourth, the fact that experts showed higher accuracy than students suggests that they performed better because they were able to overcome their initial heuristic reasoning by reasoning analytically, rather than by applying a correct heuristic. However, the reaction time outcomes from Study 3 indicate that this took more time. Fifth, our results suggest that Tversky (1997) was right concerning the vertical preference of people's perception: The effect of item type on both accuracy and reaction time of correct responses shows that most heuristic errors were made, both by students and experts, when only the height of the bars in the histograms differed. Students solved these items very fast (±2,800 ms), while experts needed more time (±4,200 ms) not to fall for the height heuristic. This suggests that Tversky's design principle is a valuable principle in the design of graphs. In order to confirm this even more, it would be interesting to do the same study with histograms that are rotated 90 deg, resulting in horizontal bars. This effect of item type on accuracy and reaction time is not only evidence for the validity of the graph design principles but can also be accounted for by dual-process theories: When solving items in which only the height differs, the lowest accuracy rates are observed as the height difference is very salient, while in other item types, the difference in position can be more easily taken into account as well.

The presented studies have several limitations. First, the rather artificial nature of the items could be seen as problematic. We do not know if participants would show the same misinterpretation and demonstrate it to the same extent in more real-life items with histograms with less "smooth" and symmetric shapes or in more ecologically valid settings such as when reading a research article. Also, results could have been slightly different when placing the to be compared histograms next to each other instead of on top of each other. However, we feel that this first study on the heuristic interpretation of histograms is a good starting point for investigating more systematically the interpretation of histograms by various groups of people using more ecologically valid items. A second limitation is that our studies only focused on accuracy rates and reaction times, while no process data were gathered. In future research, thinking aloud methods could be used to gain a deeper understanding of students' and experts' interpretation difficulties and the interplay between heuristic and analytic reasoning. Third, one could question whether all participants in Study 3 can be called true expert users of histograms, or even whether expert users of histograms really exist. In our study, we gave a rather pragmatic interpretation of the term expert: people who can be expected to be able to interpret histograms correctly, given the nature of their job or advanced study curriculum. Fourth, one might argue that, especially for experts such as the participants of Study 3, the observed differences in accuracy and especially in reaction time between congruent and incongruent items are so small that they can hardly be considered problematic from a practical point of view. However, the major point in finding the expected reaction time differences is that even experts are still affected by the height heuristic, making them in certain circumstances—e.g., when being somewhat inattentive or unmotivated—still susceptible to committing heuristic

errors. In some cases, this could even lead to an incorrect interpretation, as demonstrated in Study 3.

We now turn to some more general theoretical and methodological implications of our three studies. Afterwards, we will discuss the implications for scientific practice. Finally, we will discuss the educational implications of the results.

Our results have three important theoretical and methodological implications. First, with respect to dual-process theories, we can conclude that the dual-process theoretical framework and the method of comparing performance and reaction times on congruent and incongruent items proved to be very valuable in studying this misinterpretation of histograms. This implies that the framework and methodology may also be applicable to other graphical displays such as line graphs or bar charts. Second, we have successfully applied the graph design principles to histograms of Tversky (1997): as height is not used naturally in histograms, misinterpretations related to height occur. It is likely that other graphical displays could also be studied using these principles, possibly leading to additional suggestions to optimize their design. Recently, these design principles were also applied to box plots (Lem et al. 2013b). Third, the combination of dual-process theories and graph design principles proved to be very fruitful to study the misinterpretation of histograms: While Tversky's design principles give an account for which element of histograms will be processed first, dual-process theories allow one to explain how this element is processed and influences our reasoning. This combination of theoretical frameworks could be applied to other misinterpretations of graphical displays as well, leading to a better understanding of our interpretation of graphical displays and possibly to additional suggestions to optimize their design.

With respect to scientific practice, especially Study 3 has an important implication. As we have shown that even expert users of histograms can be misled by certain perceptual characteristics of histograms, the frequent use of histograms in scientific reports can be deemed questionable. The use of histograms could lead to incorrect interpretations of research results, with all the consequences this would entail. However, a good alternative, without any possible misinterpretations, is difficult to find. A recent study showed that the interpretation of box plots is complicated by similar factors, so they cannot be seen as a good alternative (Lem et al. 2013b).

With respect to the role of histograms in the (statistics) curriculum, our findings can provide a caveat: It is clear that histograms cannot be treated as a self-evident representation that needs little explanation. Even with warning and without time pressure, students can easily fall prey to the height heuristic, as documented in Study 2. This does not only have implications for mathematics and statistics education but also for other subject domains in which data displays play an important role, such as social sciences, geography, and biology. After all, when histograms are used to teach a certain subject, and students are not able to interpret these representations correctly, the contents to be taught are likely to be misinterpreted and learned incorrectly too. Furthermore, part of our participants in Study 3 were statistics teachers at the university level, but they still showed signs of incorrect heuristic reasoning. Further research could investigate whether statistics teachers—who should be able to diagnose and address potential misinterpretations for their students—are aware of the difficulties of histograms. This last factor, knowledge of difficulties of a certain subject, is an important part of "pedagogical content knowledge" (Shulman 2000).

More research seems to be necessary to find ways of improving the interpretation of histograms and improving the graphical design of histograms. A possible line of research could go in the direction of using multiple external representations (e.g., Ainsworth 2011) to teach histograms. Using multiple representations can help to intercept misinterpretations as the correct interpretation can be derived from other representations like a dot plot. Keller and Hirsch (1998), for example, explain that multiple external representations provide "multiple

concretizations of a concept, selectively emphasizing and de-emphasizing different aspects of complex concepts, and facilitating cognitive linking of representations thereby creating a whole that is more than the sum of its parts" (p. 1). Another possibility is to test the use of different designs of histograms like proposed by Tufte (1983) in order to make the vertical dimension less salient. When the height is made less salient, it is less likely that it will be the first element processed, making it more likely that the overall shape and horizontal position of the bars will be processed correctly. Although Tufte recommended these alternative designs, empirical research on the interpretation of these alternatives and on whether they really lead to an improvement is necessary.

**Acknowledgments** Stephanie Lem holds a PhD fellowship of the Research Foundation – Flanders (Fonds Wetenschappelijk Onderzoek – Vlaanderen). This research was partially supported by grant GOA 2006/01 "Developing adaptive expertise in mathematics education" from the Research Fund KU Leuven, Belgium.

## References

Ainsworth, S. (2011, August). *Understanding and transforming multi-representational learning.* Paper presented at the 14th European Association for Research on Learning and Instruction Conference, Exeter, UK.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2002, April). *The resilience of overgeneralization of knowledge about data representations.* Paper presented at the American Educational Research Association Conference, New Orleans, LA.

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2), 1–13.

delMas, R., Garfield, J., & Ooms, A. (2005, July). *Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions.* Paper presented at the Fourth Forum on Statistical Reasoning, Thinking, and Literacy, Auckland, New Zealand.

Gillard, E., Van Dooren, W., Schaeken, W., & Verschaffel, L. (2009). Dual–processes in the psychology of mathematics education and cognitive psychology. *Human Development, 52*(2), 95–108. doi:10.1159/000202728.

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition, 17*, 627–638. doi:10.3758/BF03197085.

Inglis, M., & Alcock, L. (2011). *Expert/novice differences in the reading of mathematics proof.* Paper presented at the 14th European Association for Research on Learning and Instruction Conference, Exeter, UK.

Inglis, M., & Simpson, A. (2004). Mathematicians and the selection task. In M. Johnsen Hoines, & A. B. Fuglestad (Eds.), *Proceedings of the 28th International Conference on the Psychology of Mathematics Education* (vol. 3, pp. 89–96), Bergen, Norway.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454. doi:10.1016/0010-0285(72)90016-3.

Keller, B. A., & Hirsch, C. R. (1998). Student preferences for representations of functions. *International Journal of Mathematical Education in Science and Technology, 29*, 1–17. doi:10.1080/0020739980290101.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2012). On the misinterpretation of histograms and box plots. *Educational Psychology [online]*. doi:10.1080/01443410.2012.674006.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013a). External representations for data distributions: In search of cognitive fit. *Statistics Education Research Journal, 12*(1), 4–19.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013b). The heuristic interpretation of box plots. *Learning and Instruction, 26*(4), 22–35. doi:10.1016/j.learninstruc.2013.01.001.

Rabinowitz, M., & Hogan, T. M. (2002). Using a triad judgment task to examine the effect of experience on problem representation in statistics. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 826–830). Hillsdale, NJ: Lawrence Erlbaum.

Rabinowitz, M., & Hogan, T. M. (2008). Experience and problem representation in statistics. *American Journal of Psychology, 121*, 395–407.

Shulman, L. S. (2000). Teacher development roles of domain expertise and pedagogical knowledge. *Journal of Applied Developmental Psychology, 21*, 129–135.

St. Evans, J. B. T. (2006). The heuristic–analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*, 378–395. doi:10.3758/BF03193858.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire: Graphic Press.
Tversky, B. (1997). *Cognitive principles of graphic displays.* Paper presented at the Association for the Advancement of Artificial Intelligence Workshop on Diagrammatic Reasoning. Cambridge, MA.
Vamvakoussi, X., Van Dooren, W., & Verschaffel, L. (2013). Educated adults are still affected by intuitions about the effect of arithmetical operations: Evidence from a reaction-time study. *Educational Studies in Mathematics, 82*, 323–330. doi:10.1007/s10649-012-9432-9438.
Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145–168. doi:10.1023/A:1003594832397.

**Stephanie Lem.** Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2 PO Box 3773, 3000 Leuven, Belgium

*Current themes of research:*

Problem solving. Statistics education. Graphical representations. Dual-process theories of reasoning.

*Most relevant publications in the field of Psychology of Education:*

Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction, 26*(4), 22–35.
Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). On the misinterpretation of histograms and box plots. *Educational Psychology: An International Journal of Experimental Educational Psychology, 33*(2), 155–174.
Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). External representations for data distributions: in search of cognitive fit. *Statistics Education Research Journal, 12*(1), 4–19.
Lem, S., Van Dooren, W., Gillard, E., Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica, 53*(2), 123–135.
Lem, S., Van Dessel, K., Vanhoof, S., Onghena, P. (2011). Attitudes toward statistics: How do they evolve during students' curriculum and what is the relation with students' evaluation of their course?. *Mediterranean Journal for Research in Mathematics Education, 10*(1–2), 43–60.

**Patrick Onghena.** Methodology of Educational Sciences Research Group, KU Leuven, Dekenstraat 2 PO Box 3700, 3000 Leuven, Belgium

*Current themes of research:*

Statistics education. $N=1$ experiments. Mixed methods research. Nonparametric statistics.

*Most relevant publications in the field of Psychology of Education:*

Luwel, K., Foustana, A., Onghena, P., Verschaffel, L. (2013). The role of verbal and performance intelligence in children's strategy selection and execution. *Learning & Individual Differences, 24*, 134–138.
Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction, 26*(4), 22–35.
Schillemans, V., Luwel, K., Onghena, P., Verschaffel, L. (2011). The influence of the previous strategy on individuals' strategy choices. *Studia Psychologica, 53*(4), 339–350.
Castro Sotos, A., Vanhoof, S., Van Den Noortgate, W., Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98–113.
Gielen, S., Peeters, E., Dochy, F., Onghena, P., Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

**Lieven Verschaffel.** Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2 PO Box 3773, 3000 Leuven, Belgium

*Current themes of research:*

Problem solving. Mathematics education. Learning and instruction.

*Most relevant publications in the field of Psychology of Education:*

Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction, 26*(4), 22–35.

Dewolf, T., Van Dooren, W., Verschaffel, L. (2011). Upper elementary school children's understanding and solution of a quantitative problem inside and outside the mathematics class. *Learning and Instruction, 21*(6), 770–780.

Fernández, C., Llinares, S., Van Dooren, W., De Bock, D., Verschaffel, L. (2012). The development of students' use of additive and proportional methods along primary and secondary school. *European Journal of Psychology of Education, 27*(3), 421–438.

De Smedt, B., Ansari, D., Grabner, R., Hannula-Sormunen, M., Schneider, M., Verschaffel, L. (2011). Cognitive neuroscience meets mathematics education: It takes two to tango. *Educational Research Review, 6*(3), 232–237.

De Smedt, B., Torbeyns, J., Stassens, N., Ghesquière, P., Verschaffel, L. (2010). Frequency, efficiency and flexibility of indirect addition in two learning environments. *Learning and Instruction, 20*(3), 205–215.

**Wim Van Dooren.** Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2 PO Box 3773, 3000 Leuven, Belgium

*Current themes of research:*

Problem solving. Mathematics education. The use of representations in mathematical problem solving. Intuitionsand biases in human reasoning. Conceptual change.

*Most relevant publications in the field of Psychology of Education:*

Lem, S., Onghena, P., Verschaffel, L., Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning fand Instruction, 26*(4), 22–35.

Obersteiner, A., Van Dooren, W., Van Hoof, J., Verschaffel, L. (2013). The natural number bias and magnitude representation in fraction comparison by expert mathematicians. *Learning and Instruction, 28*, 64–72.

Vamvakoussi, X., Van Dooren, W., Verschaffel, L. (2013). Educated adults are still affected by intuitions about the effect of arithmetical operations: evidence from a reaction-time study. *Educational Studies in Mathematics, 82*(2), 323–330.

Acevedo Nistal, A., Van Dooren, W., Verschaffel, L. (2012). What counts as a flexible representational choice? An evaluation of students' representational choices to solve linear function problems. *Instructional Science, 40*(6), 999–1019.

Van Hoof, J., Lijnen, T., Verschaffel, L., Van Dooren, W. (2013). Are secondary school students still hampered by the natural number bias? A reaction time study on fraction comparison tasks. *Research in Mathematics Education, 15*(2), 154–164